BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 6: e90987
doi: 10.3897/biss.6.90987

OPEN ACCESS

Conference Abstract

Human and Machine Working Together towards High Quality Specimen Data: Annotation and Curation of the Digital Specimen

Sam Leeflang[‡,§], Wouter Addink[‡,§], Soulaine Theocharides[‡,§]

‡ Naturalis Biodiversity Center, Leiden, Netherlands
§ Distributed System of Scientific Collections - DiSSCo, Leiden, Netherlands

Corresponding author: Sam Leeflang (sam.leeflang@naturalis.nl)

Received: 28 Jul 2022 | Published: 01 Aug 2022

Citation: Leeflang S, Addink W, Theocharides S (2022) Human and Machine Working Together towards High Quality Specimen Data: Annotation and Curation of the Digital Specimen. Biodiversity Information Science and Standards 6: e90987. https://doi.org/10.3897/biss.6.90987

Abstract

The engine for our Distributed System of Scientific Collections (DiSSCo) is running! Core technical components supporting this new research infrastructure are currently being implemented and the engine that will support it is already working. Even though some nuts and bolts may still be missing, we aim to show it in action to present how it will enable annotation and curation of the Digital Specimen. The Digital Specimen is a technical implementation based on FAIR Digital Objects (FAIR stands for Findable, Accessible, Interoperable and Reusable) to support the Digital Extended Specimen concept (Webster et al. 2021). We will also present and demonstrate how we will implement standardized quality checks as they are being developed in Biodiversity Information Standards (TDWG) to enhance the quality of the data.

DiSSCo is currently in its preparation phase. This phase will end in January 2023 with the completion of the DiSSCo Prepare project funded by the European Commission. Part of that project is the design of the Digital Specimen infrastructure, which is not an easy task considering the wide range of use cases, stakeholders and the many possibilities it offers. However, as we are moving towards the end of the project, we have defined clear goals and priorities to give shape to that infrastructure. This is where we take a fail fast approach: to quickly implement the proposed solution and see if it really fits.

One of the major needs we want to support with the Digital Specimen infrastructure (based on collected user stories (Fitzgerald et al. 2021)) is to provide services for improving the quality and usability of specimen data. Our infrastructure aims to support annotating and community-curation of the data by both machines and users. Examples of these annotations are image-based determinations, automated- or citizen science-contributed label translations or the semi-automated linking with other biodiversity data. Semi-automated linking is currently being piloted as part of the Biodiversity Community Integrated Knowledge Library project (BiCIKL) and will use a process of link prediction through artificial intelligence in combination with human validation. Improvements in data quality made together by human and machine through the curation and annotation services will help in producing a digital specimen data object with high quality, curated and extended data.

As part of the presentation we aim to give a live demonstration with the first setup in which we will ingest a dataset, run standardized quality checks and automated data enrichment services. The end result will be a digital specimen that we will present in a user-friendly interface, which has been validated by quality checks and annotated by both a human and a machine. The result will also be accessible as a FAIR Digital Object through an API. During the demonstration, we aim to give the audience a clear view on how DiSSCo can help them create higher quality specimen data, and how we will benefit in this process from the outputs of the TDWG Data quality tests and assertions taskgroup.

Keywords

data enrichment, annotation, data curation, FAIR, digital object, DiSSCo, digital specimen, DiSSCo Prepare, data quality, BiCIKL, infrastructure

Presenting author

Sam Leeflang

Presented at

TDWG 2022

References

- Fitzgerald H, Juslén A, von Mering S, Petersen M, Raes N, Islam S, Berger F, von Bonsdorff T, Figueira R, Haston E, Häffner E, Livermore L, Runnel V, De Smedt S, Vincent S, Weiland C (2021) DiSSCo Prepare Deliverable D1.1 Report on Life sciences use cases and user stories. https://doi.org/10.34960/xhxw-cb79

- Webster M, Buschbom J, Hardisty A, Bentley A (2021) The Digital Extended Specimen will Enable New Science and Applications. Biodiversity Information Science and Standards 5 https://doi.org/10.3897/biss.5.75736